e. Are specific quality descriptors of text-based reviews such as ‘enthusiastic’, ‘disappointed’, and others, strongly associated with rating levels?

Considering this question, we are required to select specific quality descriptors of text-based reviews to analyze the association between these words and rating levels. Hence, the first job we have to do is to select specific words. And then we plan to regard these words as features each of which denotes the frequency of the word in a sentence. In the last step, we plan to analyze the correlation between tf-idf of every word and rating levels.

1.Selecting quality descriptors

1.Tokenization

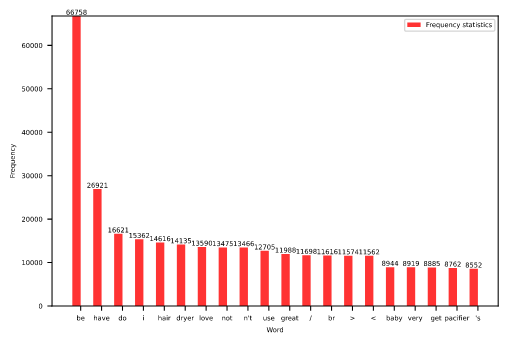
Because we want to analyze the correlation between frequency of every select quality descriptor we choose and rating levels, we plan to split the sentences into words and to count the number of each words in order to select those most frequent words that can serve as quality descriptors. Therefore, we use words tokenization that based on the space between every word to select sentences.

2. Lemmatization

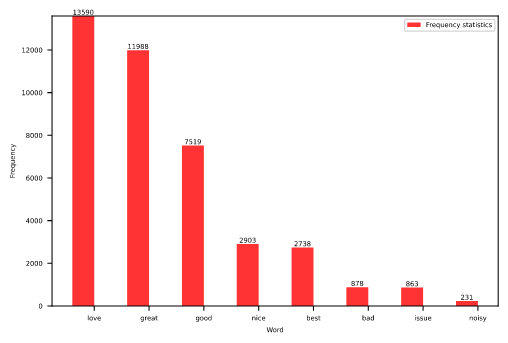
Besides, considering some words have different tenses or parts of speech, different tenses or parts of speech of the word needs to be served as the same word in order to better calculate every words’ frequency. Therefore, we need to group together them into their lemma based on its intended meaning. NLTK provides tagging and lemmatization methods that are able to precisely determining the lemma of words based on their potential meaning.

3.Counting and selection

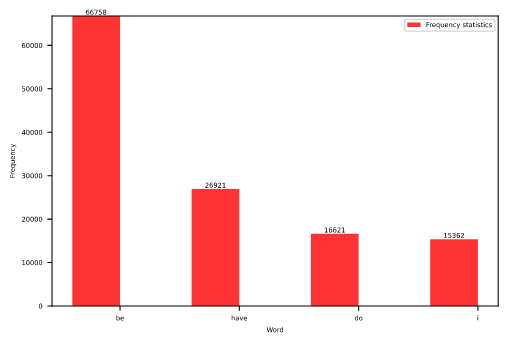
We exclude function words which generally cannot be quality descriptors. As a result, we get the word list and plot the most common 20 words’ frequency as follow.



These are the most frequent words. However, not all of these can serve as quality descriptors or even valid words such as ‘<’. We choose words that can act as quality descriptors as follow. And there shows the frequency as well.



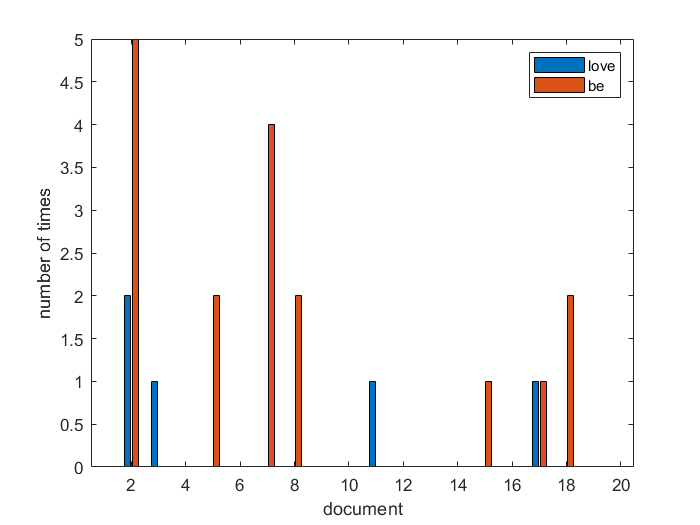
In order to directly recognize the possible correlation, we select some neutral words as control group.

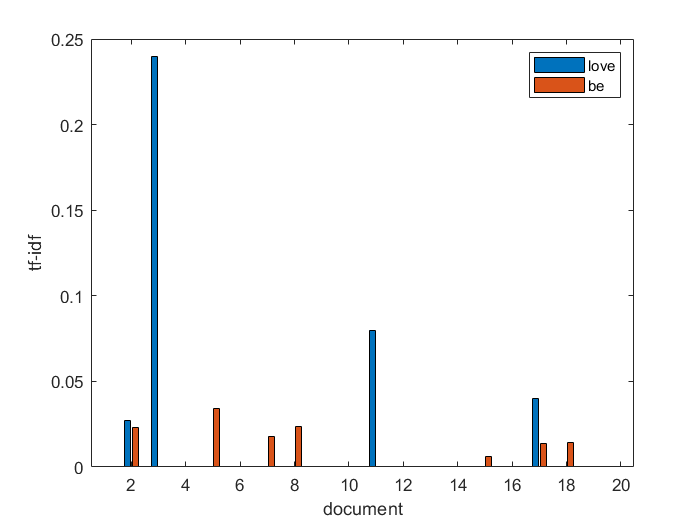
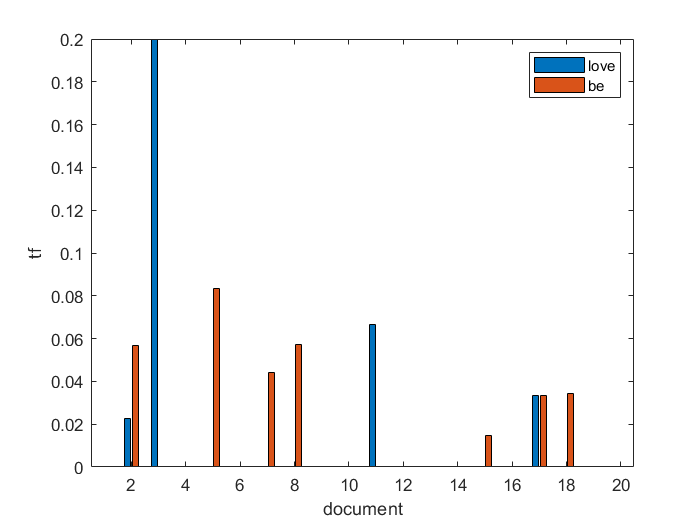


2.Correlation analysis

1,Tf-idf

Tf-idf stands for term frequency–inverse document frequency. It bases on an assumption that the most valuable words should be those words that are most frequent in the specific document and the least frequent in all the documents at the same time. Since the proportion of every word is not even as shown above, we choose to use Tf-idf of those quality descriptors as features.





2. Correlation analysis

用2个指标两两间的线性相关系数来衡量，即R统计量进行计算，



表示正线性相关， 表示负线性相关， 在0附近表示两随机变量基本上不相关。

The result is as follow.

love great good nice best bad issue noisy be

0.168 0.127 0.042 0.034 0.077 -0.075 -0.01 -0.014 -0.060

have do i

-0.014 -0.157 -0.031

We find that all of these words’ correlation coefficient are not high enough. However, some of these quality descriptors’ correlation coefficient are higher than others’. Then we find that these words are not necessary to a positive or negative review. Therefore, we combine some of these words and calculate the correlation coefficient again. The result is as follow.

love , great, good, nice, best bad, issue, noisy

0.215 -0.113

The result is better. Hence, we draw the conclusion that they are associated with star ranking.